A Formal Analysis of Classical Greek Conjugation

Ethan Yates
Brandeis University
Waltham, MA
ethanyates@brandeis.edu

Abstract

This article is an attempt to formalize the Greek conjugational system with mathematical notation as a means of working towards a computational implementation of the synthesis of verb forms. It is hoped that such a tool will be of use to students who are learning the nuances of the complex verbal system in classical Greek. With this attempt at formalization, it is hoped that it will provide the core basis for the development of a morphologically intelligent information retrieval system for classical Greek utiziling modern analyses and tools.

1 Introduction

Classical Greek has been studied extensively for thousands of years. Reading knowledge of classical Greek (henceforth referred to just as *Greek*) is a sought-after skill in the humanities, especially in the fields of classical studies, theology, philosophy, and linguistics. As a result of its long history, the grammar of classical Greek has been very thoroughly studied. However, the main practitioners of classical Greek work outside of the field of linguistics (classical studies, theology, and philosophy) meaning the amount the most modern methods in linguistics have yet to be used to analyze the language to the same degree as modern languages.

One of the most important resources for studying Greek is the *Perseus Digital Library*, a platform that hosts ancient texts (in both Latin and Greek) in a reading environment. The platform allows users to click on individual words in order to show their lemma and morphological analysis.

Morphologically intelligent information retrieval (expecially verbs) is of utmost importance to students of classical Greek (hereafter referred to just as *Greek*). Verb conjugations are notoriously difficult for students in their first year of study, due to the relatively high complexity of the conjugational system in comparison to modern languages more widely spoken Europe and North America with simpler verbal paradigms. The conjugational system utilizes thematic vowels, concatenation of personal endings, reduplication, and derivational prefixes, resulting in around 2 million potential forms for each verb (Crane, 1991). Most textbooks and reference grammars have examples of

There is in fact already a program that does this–*Morpheus* is a program created by Perseus Digital Library to take Greek words as input and output their morphological analysis. Morpheus was used to create the treebanks of Greek texts hosted by Perseus. These treebanks (organized by individual text) contain texts organized by sentence, with each word tagged for its morphology and dependency syntax. While they have quite extensive coverage of the most famous Greek texts, not every text has been analyzed. Classical Greek, having a written history spanning thousands of years, has many texts which have not been digitzed from old printed editions, and even many texts that have never been edited into print from the original manuscripts.

Morpheus was written in the C programming language decades ago, using a very rigorous set of rules written by classical scholars. A new program written in a language more suitable for data manipulation (such as Python) is desireable. The amount of an-

notated data in the treebanks is enough to work towards building a model using statistical methods. However, given the relatively large amount of features for each verb form, it is necessary to first formulate a computational analysis of Greek conjugation to allow for the generation of verb forms, given a lemma and a set of features as input. Once it is possible to generate verb forms, it is an important step towards building a system that works in the *opposite* direction, to parse the morphological features and the lemma from a given verb form. This paper will focus on building a formal analysis of Greek conjugation (inspired by previous work on other languages) as a first step toward building a program that will synthesize verb forms and provide a basis for a larger system. In addition, it will be demonstrated how such synthesis of verb forms can be used to build corpora for statistical analysis.

2 Prior work on Conjugation

2.1 Linguistic treatments of conjugation

Greek grammar has a very long tradition stretching back to antiquity. The *modern* study of Greek grammar dates back c. 200 years. In this computation analysis, it is desireable to use the most thorough analysis of the conjugational system as possible, while drawing potential insights from older sources. In the English language, the pioneering reference grammar is *A Greek Grammar for Colleges* (Smyth, 1920) which describes the entire language in detail. Almost a century later, Cambridge University published *The Cambridge Grammar of Classical Greek* (van Emde Boas et al., 2019), which represents the most "up to date" reference Grammar published in English. The system presented in this paper is based on the sections on verb inflection in both reference grammars.

Both reference grammars provide a mostly *synchronic* analysis (dipping into diachronic analysis in some areas to give context). Given that historical linguistics often provides explanation for so-called "irregular" forms (in traditional classroom pedagogy), it was necessary to consult *The Origins of the Greek Verb* (Willi, 2018), which provides the most thorough analysis of Greek conjugation in the context of Indo-European historical linguistics.

In 1948, Roman Jakobson published a famous alaysis of Russian conjugation in which he breaks the verbal system down into a system of rules (in prose form) (Jakobson, 1948). This approach was well recieved and emulated by other authors in analyses of Polish (Schenker, 1951), Czech (Rubenstein, 1951), and Old Church Slavonic/Old Russian (Halle, 1951).

2.2 Computational treatments of conjugation

Multiple scholars were inspired by Jakobson's rule-based approach. A computational implementation of Jakobson (1948) was written in the ALGOL60 programming language (Kortlant, 1971). Mathematician Joachim Lambek, noticing how a rule-based system lends itself to a mathematical analysis, wrote a computational analysis of French conjugation (Lambek, 1975). The main development by Lambek was the formalization of a rule-based system like in Jakobson (1948), into rewrite rules, lending themselves to sting manipulation.

As mentioned before, *Morpheus* was developed for morphologically intelligent information retrieval utilizing an extensive rule–based system curated by experts (Crane, 1991). This program both parses and generates forms of words, and has been used to generate large annotated corpora. These rules are extremely numerous, and designed for multiple dialects. The program is quite old (written in *C*) and does not utilize statistical methods. Working towards a new program for morphologically intelligent information retrieval is the main motivation for this paper.

3 Scope of analysis

The full paradigm of a Greek verb contains up to 1000 distinct forms, and when including

derivational prefixes, can number in the millions (Crane, 1991). Therefore, it is necessary to lessen the scope of the analyses of verbs. In this article, only the *indicative* voice of the *regular* verb (terms to be explained later) will be analyzed in order to create the beginnings of a system that can be expanded to cover all types of Greek verbs. This paper will first give as concise a *linguistic* description of conjugation as is possible, followed by a *mathematical* analysis of the conjugational system. The end of the article will contain some proposals as to how the data from this program could be utilized.

4 Linguistic overview of Greek conjugation

4.1 Features of Verbs

When it comes to morphologically intelligent information retrieval, it is necessary to define what information is being retrieved (or in the opposite case of the synthesis of verb forms, what information is given to the function as input). There are **five** features to be extracted. Verbs are marked for *person*, *number*, *tense*, *voice*, and *mood*. Each potential option for these features is given below.

- **Person**: *first, second, third*
- **Number**: *singular, dual, plural*
- **Tense**: *present, imperfect, future, aorist, perfect, pluperfect.*
- **Voice**: *active, middle, passive*
- **Mood**: *indicative, subjunctive, optative*

Not all tense-mood combinations exist–Table 1 shows which combinations have conjugations. Traditional pedagogy designates a *sequence of tenses*, dividing tenses into *primary* and *secondary* tenses (in modern linguistic literature called *non-past* and *past*, such as in Willi (2018)). This classification determines which prefixes and personal endings are used when selecting morphemes. These "tenses" are in fact **tense-aspect** combinations, and a tense-aspect combination will be referred to as a **tense-system**.

At this stage, only paradigms in the active voice and indicative mood will be considered, for the sake of keeping the scope reasonable and for the fact that *all* tense systems are represented within the paradigms of the indicative mood.

4.2 Tense and Aspect

While Greek is usually described as having six "tenses", these "tenses" are more accurately described as morphological systems designated to a tense-aspect pair. Tenses are divided into two "sequences" (tenses). There are also three apects–*perfective, imperfective*, and *stative* (Willi, 2018). The six systems (referred to as "tenses") are yielded by the combination of two sequences (*non-past* and *past*) with three aspects. The tense aspectual pairs are shown in Table 2.

This table will be relevant during formalization, because it represents the mapping of tense-systems to tense-aspect pairs. As will be seen in the analysis of an example paradigm, some morphemes are conditioned on the tense (non-past or past) or the aspect, as opposed to the tense system itself (present, future, perfect, etc.). Again, what are referred to as "tenses" in the Greek grammatical tradition, are actual tense-apsect pairs, and thus it is preferable for the formalization to utilize the traditional terminology in the input, map the "tense" to a tense-aspect pair, and then use those features for the selection of morphemes.

4.3 Romanization

This paper will require a consistent system of Romanization to be accessible to those without knowledge of the Greek writing system. The table of romanization is given in Tabel 3.

This particular system of Romanization was adapted from that used by Wiktionary in all of their entries for Greek. [1] Wiktionary describes their system as *scientific transliteration*–it is desireable because it is designed for automatic computation, but is also readable for scholars when examining Greek entries.

[1] https://en.wiktionary.org/wiki/Wiktionary:Ancient_Greek_transliteration

	Indicative	Subjunctive	Optative	Imperative	Sequence
Present	+	+	+	+	**primary**
Imperfect	+	-	-	-	**secondary**
Future	+	-	+	-	**primary**
Aorist	+	+	+	+	**secondary**
Perfect	+	+	+	+	**primary**
Pluperfect	+	-	-	-	**secondary**

Table 1: Tense-Mood Paragigms

	Perfective	Imperfective	Stative
non-past	Future	Present	Perfect
past	Aorist	Imperfect	Pluperfect

Table 2: Table of the tense-aspect system

Greek	Latin
α	a
β	b
γ	g
δ	d
ε	e
ζ	z
η	ē
θ	t̲
ι	i
κ	k
λ	l
μ	m
ν	n
ξ	x
ο	o
π	p
ρ	r
σ	s
τ	t
υ	u
φ	p̄
χ	k̲
ψ	p̲
ω	ō

Table 3: Romanization scheme

Their system uses several digraphs (*ph, th, kh, ps, ks, rh*) to represent what were originally single graphemes. This is not a problem when used as a one-way transliteration system for scholarly reference. But for computation, it is a lot less confusing if there is a one-to-one correspondence between each Greek letter and its romanization. For this project, several modifications have been made to Wiktionary's system in order to remove all digraphs: **ph**, **th**, **kh**, **ks**, **rh** are changed to **p̄**, **t̲**, **k̲**, **x**, **r**. These representations are inspired from the scholarly transliteration of Arabic from Hans Wehr's Arabic Dictionary (Wehr, 1961). This system was chosen in order to keep the three way distinction between voiced, aspirated, and unaspirated stops.

One final digraph not mentioned is the Greek letter Psi **ψ**, usually romanized as **ps**. In order to mark that it is one grapheme instead of two, it is written as p̲ (*p* with a cedilla diacritic beow the letter, representing a little *s*).

In terms of vowels there are the typical five vowels–*a, e, i, o, u* along with their accompanying long counterparts ($\bar{a}$,$\bar{e}$,$\bar{\imath}$,$\bar{o}$,$\bar{u}$).

This scheme of romanization is provisional, and only used for the purposes of this paper. The use of diacritics is meant to reduce the amount of graphemes, so that each grapheme in Greek is represented by exactly one grapheme in romanization.

Greek has a complex system of accentuation. In the scope of this paper, this will not be addressed–this is fine for our purposes, and has been an approach taken in prior projects of a similar nature (Crane, 1991). However, it will be a necessary step to address it in future work. For the sake of completeness, accents are still included in romanization.

4.4 Description of graphemes

Graphemes will briefly be presented based on the properties of their respective phonemes. There are 9 *stops* (also called *plosives*), two nasals, a liquid, and a silibant. The 9 stops show a three-way distinction based on voicing and aspiration. The distribution of consonants is shown in Table 4.

4.5 Morphemes and Lexemes

The active-indicative conjugation of *lúō*, the archetypal regular verb, is given in table 5 (the dual number has been omitted for brevity).

4.5.1 Overview

The Greek verb consists of *at most* 5 elements (in reality 6, but here we ignore derivational morphology and focus only on inflectional morphology). Notice how in table 5 the presence of the e-augment in the *past tenses*, and its absence in the *non-past* tenses. In addition, take note of the tense-suffixes present for each aspect (-$\emptyset$- in the *perfective*, -*s*- in the *imperfective*, and -k- in the *stative*).

- 2 prefixes
    1. *e*-augment (also called *epsilon augment*)
       present-stem *lu-* > imperfect-stem *élu-*
    2. Reduplication[2] of the stem initial consonant with epenthetic *e*
       verb-stem *lu-* > reduplicated *lelu-*
- The *verb-stem* as specified in a lexicon. In this case, the archetypal verb *lu-* (”to loosen”/”to unbind”)
- 3 suffixes
    1. A *tense suffix*, determined by the aspect (perfective, imperfective, or stative)
       tense-suffix -*k*- for stative *lélu-k-a*, tense-suffix -*s*- for 1st singular aorist *élú-s-ate*
    2. A *thematic vowel*, determined by the tense, person, and number.
       Thematic vowel -*o*- in 1 singular imperfect *élu-o-n*
    3. A *personal ending*, determined by the tense, person, and nummber.
       Personal ending -*n* in *éluo-n*

For each of the four morphemes, we will define a table/matrix defining which features condition their usage, and use the defined system to describe an example paradigm. It is important to note the distinction between the *verb-stem* and the *tense-stem*. The verb-stem is the smallest unit with no added features drawn from the lexicon, while the tense-stem consists of the verb-stem combined with the prefixes (or lack thereof) and *tense-suffix* (or lack thereof) conditioned by the tense-system (tense-aspect pair). There is a minimum of three morphemes (verb stem, thematic vowel, and personal ending).

4.5.2 Prefixes

The *e*-augment is prefixed to the beginning of the verb stem in the past-tense (secondary sequence, or the imperfect, aorist, and pluperfect tenses). Reduplication of the initial consonant occurs in the stative aspect (perfect and pluperfect tenses). The following table shows their distribution.

This is where considerable economy is obtained in the conditioning of morphemes when tense and aspect are separated from the traditional *tenses*. It is clear that each morpheme is conditioned by a single feature (with overlap in the past stative). But if there was no distinction between tense and aspect,

[2] Reduplicated consonants undergo *deaspiration* if the initial consonant of the verb stem is aspireated. e.g. *ṭú-* $>^{perf.}$ *teṭuk-*

	unaspirated	aspirated	voiced	nasalized	liquid	trill	silibant
labial	p	p̄	b	m			
dental	t	t̲	d	n	l	r	s
velar	k	k̲	g				

Table 4: Distribution of cononants (graphemes)

Primary	**Secondary**
PRESENT	IMPERFECT
lúō	élуon
lúeis	élues
lúei	élue
lúomen	elúomen
lúete	elúete
lúousi	elúousi
FUTURE	AORIST
lúsō	élusa
lúseis	élusas
lúsei	éluse
lúsomen	elúsamen
lúsete	elúsate
lúsousi	elúsan
PERFECT	PLUPERFECT
léluka	elelúkein
lélukas	elelúkeis
léluke	elelúkei
lelúkamen	elelúkemen
lelúkate	elelúkete
lelúkasi	elelúkesan

Table 5: Regular verb conjugation

all of the tense-systems would require morphemes assigned to them individually.

4.5.3 Suffixes

The Greek verb has at most three suffixes.

1. The *tense-suffix* is a suffix added to the verb-stem conditioned on the aspect. The three tense-suffixes are:
   - *-s-* for the **perfective**
   - *-∅-* for the **imperfective**
   - *-k-* for the **stative**
2. The *thematic vowel* is infixed between the tense-stem and the personal ending. There are three variations of the thematic vowel that are conditioned on the tense-system; *e/o*, *a*, and *e*. The distribution of these theme vowels is conditioned by the tense-system and the person-number, and is shown below in Table 7.
3. The *personal ending* is conditioned by the person and number, and is the final prefix. There are 5 different sets of endings, but we will only focus on the *thematic* endings, both *primary* and *secondary*. All the endings will be given for the sake of completeness, but we are not addressing the middle-passive system in this paper. These are given in table 8

All of these suffixes can be represented as matrices, the dimensions of which correspond to each feature conditioning each respective suffix. This will be defined explicitely in the next section.

5 Mathematical analysis and formalization

Now that the conjugation of the regular verb (with our subset of restrictions) has been de-

	Perfective	Imperfective	Stative
non-past	∅	∅	reduplication
past	*e*-augment	*e*-augment	*e*-augment+reduplication

Table 6: Distribution of prefixes

	1s	2s	3s	1p	2p	3p
Present Future Imperfect	o	e		o	e	o
Aorist Perfect	a					
Pluperfect	ei			e		

Table 7: Theme vowel matrix

scribed in linguistic terms, it is now necessary to give a mathematical and computational formulation of all these rules. The following mathematical analysis is primarily inspired by the approach in Lambek (1975). Each set of morphemes can be viewed as an tuple, conditioned by cardinality (or indices).

We will first need to define a tuple containg tenses.

$$i = \{nonpast, past\}$$

In tuple j, we list all three aspects. As above, these are accessed via index, or cardinality. Therefore, if $j = 1$, then $j = perfective$.

$$j = \{perfective, imperfective, stative\}$$

Traditionally, tense-aspect combinations are not referred to by a combination of the names of each respective tense and aspect. Therefore, we define a tuple k containing the names of each tense-aspect pair. We will provide a mapping between i, j and k after defining the personal endings.

$$k = \begin{array}{c} \{future,\ present,\ perfect, \\ aorist,\ imperfect,\ pluperfect\} \end{array}$$

There are 6 possible personal endings. Technically there are 8, but the dual number (which is not even present in many dialects, and never present in the 1st person) has been excluded in order to reduce the scope of this paper.

$$l = \{1s, 2s, 3s, 1p, 2p, 3p\}$$

Here, a mapping is given for mapping each *tense system* to the respective tense-aspect pair.

$$k = \begin{cases} 1,1 \text{ if } k = 1 \\ 1,2 \text{ if } k = 2 \\ 1,3 \text{ if } k = 3 \\ 2,1 \text{ if } k = 4 \\ 2,2 \text{ if } k = 5 \\ 2,3 \text{ if } k = 6 \end{cases}$$

The last variable we need to define is the **stem**. Let S be a tuple, representing an array of graphemes in the stem. Thus, the graphemes in the stem S can be accessed with an index. For example, if $S = lu$-, then $S = \{l, u\}$m and $S_1 = l$

This brings us to the main conjugation function C which takes only a stem as input.

$$C_{k,l}(S) = \textsc{Prefix}^S_{i,j} + S + \textsc{Suffix}_{k,l} \quad (1)$$

The function for generating the prefix takes the stem S and the tense aspect pair (i, j) as input. We let ρ represent the reduplication prefix, and α as the e-augmentation prefix.

$$\textsc{Prefix}^S_{i,j} = \alpha_i + \rho_{S,i} \quad (2)$$

The conditioning of reduplication ρ is as follows.

$$\rho_{S,j} = \begin{cases} S_1\text{e if } j = 3 \\ \emptyset \textit{ otherwise} \end{cases} \quad (3)$$

	Active			Middle-Passive	
	Primary		Secondary	Primary	Secondary
	Thematic	Athematic			
1sg	–	-mi	-n, –	-mai	-mēn
2sg	-si	-s	-s	-sai	-so
3sg	-ti	-si(n)	–	-tai	-to
1pl	-men	-men	-men	-meṭa	-meṭa
2pl	-te	-te	-te	-sṭe	-sṭe
3pl	-nsi	-nsi	-n, -san	-ntai	-nto

Table 8: Personal endings

The conditioning of the e-augment is as follows.

$$\alpha_i = \begin{cases} \dot{\text{e}} \text{ if } i = 2 \\ \emptyset \text{ otherwise} \end{cases} \tag{4}$$

Finally, the suffixes are obtained from selection of the theme vowel and personal ending from their respective matrices. These matrices are represented by tables 7 and 8, respectively.

$$\textsc{Suffix}_{k,l} = \textsc{Theme}_{k,l} + \textsc{Ending}_{i,l} \tag{5}$$

6 Uses of Current Formalization

Here is an example derivation, showing the synthesis of the 3rd person singular, present active indicative form of the archetype verb *lúō* "to unbind, to loosen".

$$C_{k,l}(S) = \textsc{Prefix}^{S}_{i,j} + S + \textsc{Suffix}_{k,l}$$

Here we substitute the parameters (stem and morphological features) into their respective variable.

$$\textsc{Prefix}^{\text{lu}}_{1,1} + \text{lu} + \textsc{Suffix}_{2,3}$$

$$\alpha_1 \rho_{\text{lu},1} \text{luSuffix}_{2,3}$$

$$\emptyset\emptyset\text{luSuffix}_{2,3}$$

When there are *zeros*, they can be eliminated.

$$\text{luSuffix}_{2,3}$$

$$\text{luTheme}_{2,3}\text{Ending}_{1,3}$$

$$\text{lueEnding}_{1,3}$$

$$\text{lueti}$$

We can finally say that $C_{2,3}(\text{lu}) = \textit{lueti}$. However, there is an issue–in table 5, the 3rd person singular present active indicative form of *lúō* is *lúei*, not **lúeti*.

This discrepency is due to a historical development of Greek phonology, in which intervocalic -*t*- became -*s*- regularly. This was followed by another change, in which intervocalic -*s*- was eliminated regularly. With these two rules in mind, we can illustrate the derivation of the actual form *lúei* from the form that was synthesized by the given functions (**lúeti*).

**lúeti* > **lúesi* > **lúei*

Since these functions work on the level of *graphemes* and not *phonemes*, derivations (or alternatively called *rewrites*) are labelled as **morphographemic rules**. In philological terminology, the form generated before the application of morphographemic rules is calle the *underlying form*, while the form synthesized after the application of the morphographemic rules is called the *surface form*. In table 9, the present tense conjugation of *lúō* is given, comparing the underlying forms with the surface forms. Wherever there is a discrepency, the underlying form is marked with an asterisk.

Underlying	Surface
*luo∅	luō
*luesi	lueis
*lueti	luei
luomen	luomen
luete	luete
*luonsi	luousi

Table 9: Underlying forms compared to surface forms in the present-tense conjugation

Transforming the underlying forms into the surface forms, in the past, has meant hand-writing a very large number of rules and testsing for their accuracy.

7 Conclusion

Using a system such as the one described in this paper, it is possible to compose a corpus of *underlying forms* aligned with *surface forms*. Using statistical methods, the underlying forms generated by conjugation function, aligned with verb forms from preexisting corpora, provides the potential for eliminating the laborious step of writing morphographemic rules, with statistical tools (such as a neural network sequence to sequence model) inferring the rules fom the given data. It seems that this is a potentially powerful line of inquiry that should be explored further.

References

Gregory Crane. 1991. Generating and parsing classical Greek. *Literary and Linguistic Computing*, 6:243–245.

Morris Halle. 1951. The Old Church Slavonic conjugation (with an appendix on the Old Russian conjugation). *Word*, 7:155–167.

R. Jakobson. 1948. Russian conjugation. *Word*, 4:155–167.

F.H.H. Kortlant. 1971. Russian conjugation: Computer synthesis of Russian verb forms. *Tijdschrift voor Slavische Taal- en Letterkunde*, 1:51–80.

J. Lambek. 1975. A mathematician looks at French conjugation. *Theoretical Linguistics*, 2:203–214.

Herbert Rubenstein. 1951. The Czech conjugation. *Word*, 7:144–154.

Alexander M. Schenker. 1951. Polish conjugation. *Word*, 10:469–481.

H.W. Smyth. 1920. *A Greek Grammar For Colleges*. Harvard University Press, Cambridge, MA.

Evert van Emde Boas, Albert Rijksbaron, Mathieu de Bakker, and Luuk Huitink. 2019. *The Cambridge Grammar of Classical Greek*. Cambridge University Press, Cambridge, UK.

Hans Wehr. 1961. *A Dictionary of Modern Written Arabic*. Otto Harrassowitz, Wiesbaden, Hesse.

Andreas Willi. 2018. *The Origins of the Greek Verb*. Cambridge University Press, Cambridge, UK.